Speech Communication 100 (2018) 41–45

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Acoustic classification of Russian plain and palatalized sibilant fricatives: Spectral vs. cepstral measures

Laura Spinu[a], Alexei Kochetov[b], Jason Lilley[c]

[a] Department of Communications & Performing Arts, CUNY Kingsborough Community College, 2001 Oriental Boulevard, Office E313 Brooklyn, NY 11235 USA
[b] Department of Linguistics, University of Toronto, Toronto, ON, Canada M5S 3G3
[c] Center for Pediatric and Auditory Speech Sciences, Nemours Biomedical Research, Wilmington, DE 19803, USA

ABSTRACT

This study compares two methods for classifying voiceless sibilant fricatives forming a 4-way phonemic contrast found in Russian, but otherwise cross-linguistically rare. One method uses spectral measures, i.e. vowel formants, COG, duration and intensity of frication. The second method uses cepstral coefficients extracted from different regions inside fricatives and neighboring vowels. The corpus comprises 1,431 plain and palatalized fricatives from two places of articulation, produced by 10 speakers. Logistic regression was used to classify the productions of males and females together and separately. The productions of females yielded higher correct classification rates (highest 91.9%). Cepstral measures outperformed spectral measures across-the-board.

1. Introduction

The acoustic characteristics of fricatives have been the focus of numerous studies. While some studies sought to identify the general properties of different fricatives cross-linguistically (Gordon et al., 2002; Tabain and Watson, 1996), others narrowed down the focus of their study to a specific language (Jongman et al., 2000; Jesus and Shadle, 2002, 2003; Maniwa et al., 2009), or even more specific aspects pertaining to fricatives within a language, such as whistled fricatives (Lee-Kim et al., 2014), laryngeal articulations (Nawrocki, 2008), pharyngeal articulation (Proctor et al., 2010), devoicing (Pape and Jesus, 2015), secondary palatalization (Spinu and Lilley, 2016; Kochetov, 2017), or sibilance (Flipsen et al., 1999).

Russian is one of the few languages with a 4-way contrast involving palatalized sibilant fricatives, specifically: palatalized dental/alveolar /sʲ/, palatalized post-alveolar (prepalatal) /ʃʲ/, non-palatalized dental/alveolar /s/ and retroflex (apical post-alveolar) /ʂ/ (Timberlake, 2004). By investigating a contrast that is noteworthy in its rare cross-linguistic occurrence, the current study adds to the growing body of work on the acoustics of fricatives. Our goal is to identify the best methodological ways to classify fricatives accurately. We thus compare the performance of two classification methods, one based on spectral measures traditionally used in phonetic research and a method based on cepstral measures, on a corpus consisting of plain and palatalized Russian voiceless sibilant fricatives.

2. Background

The most common acoustic measures previously used with fricatives include center of gravity, spectral peak location, spectral slope, spectral moments, noise duration, F2 onset frequency, static and dynamic amplitude measurements, and locus equations (see McMurray and Jongman, 2011, for a comprehensive review). While traditionally these measures were based on discrete Fourier transforms, multitaper spectra were recently introduced as better suited for stochastic parts of speech (Koenig et al., 2013; Lousada et al., 2012; Zygis et al., 2012). Much of the previous work on fricatives focused on identifying parameters that differ significantly between various categories (most commonly, place of articulation and/or voicing), but relatively few of them were designed specifically for classification, that is, in order to identify acoustic parameters able to reliably differentiate or discriminate a corpus of fricative consonants in terms of place of articulation and voicing.

Early studies performing classification of fricatives based on spectral moments yielded correct classification rates of 74.5%-77.7% (Forrest et al., 1988, with a set comprised of /f, θ, s, ʃ/) and 74%-78% (Tomiak, 1990, with /f, θ, s, ʃ, h/). Using spectral moments, duration, normalized amplitude, and spectral slope, Nissen and Fox (2005) obtained a correct classification rate of 65% for adult productions of /f, θ, s, ʃ/. One of the most comprehensive recent classification studies employed a corpus of eight fricatives at four places of articulation, specifically /f, v, θ, ð, s, z, ʃ, ʒ/, produced by 20 English speakers (McMurray and Jongman, 2011). Using 24 predictors comprising all the best known measures as well as newly-developed ones, the authors

E-mail address: lspinu@kbcc.cuny.edu (L. Spinu).

https://doi.org/10.1016/j.specom.2018.04.010
Received 29 September 2017; Received in revised form 11 April 2018; Accepted 25 April 2018
Available online 27 April 2018
0167-6393/

obtained correct classification rates between 79.2-85%, improving on Jongman et al.'s previous classification rate for the same corpus, i.e. 77% (Jongman et al., 2000).

Spinu and Lilley (2016) compared a novel method, based on cepstral coefficients, with a method based on spectral moments to classify 5 pairs of plain and palatalized Romanian fricatives (i.e. [f-fʲ, v-vʲ, z-zʲ, ʃ-ʃʲ, x/h-çʲ]), produced by 31 native speakers and obtained a correct classification rate of 95.3%. Hidden Markov Models (HMMs) were used to divide each fricative into three regions based on their internal variance. The cues were extracted from regions inside the frication portion only, without any information from adjacent vowels. Crucially, their corpus did not include interdental fricatives, which have traditionally posed challenges to classification. The classification rate obtained may thus not be a major improvement over previous studies, but rather complementary to them, contributing data from a different part of the fricative 'landscape'. Even so, the high correct classification rates obtained recommend the cepstral coefficient/HMM-region method as a reliable way of identifying the crucial combination of a fricative's properties that makes it unique and distinguishable compared to all other fricatives. Cepstral coefficients were shown to yield more accurate classification of place of articulation, palatalization, voicing and gender compared to spectral moments. No comparison was made with more traditional fricative measurements such as formants and center of gravity. The current study expands the comparison to these other measures.

3. Current study

In this study, we extend the method from Spinu and Lilley (2016) to a new language, Russian, focusing on a specific subset of fricatives – voiceless plain and palatalized sibilants. Adding to the complexity of the problem is the fact that this 4-way primary/secondary place contrast is extremely rare cross-linguistically. The absence or rarity of contrasts in languages is often associated with facts of phonetic difficulty[1] and resulting phonological instability (Hayes and Steriade, 2004). This contrast therefore constitutes a good testing ground for the HMM-region cepstral-based classification method (HCCM). Our corpus was originally collected for Kochetov (2017), where it was used to provide a general acoustic description of the Russian contrasts. The materials consisted of 48 target words with the fricatives /s/, /sʲ/, /ʂ/, and /ʃʲ/. The words were produced 3 times in a carrier phrase by 10 native speakers of Standard Russian (5 females and 5 males; median age 21.5), all born and raised in Russia, but at the time of the study residing in Canada. The resulting 1,431 tokens (144 tokens per speaker, minus 9 omissions) were annotated in Praat (Boersma and Weenink, 2015), indicating the fricatives and preceding/following vowels. Please see Kochetov (2017) for additional details regarding the participants' background, recording procedure, and annotation criteria.

3.1. Spectral measures, duration, and amplitude

For the purposes of this study, a Praat script was run on the corpus data to obtain duration plus the following 15 measurements from the annotated data:

Duration (in milliseconds) of the fricative and adjacent (word-internal) vowel

Amplitude (in dB) of the fricative and adjacent (word-internal) vowel

Centre of gravity of fricative noise (COG, or the first spectral moment, in Hz), measured at 3 points in time: onset (C-on), midpoint (C-mid), and offset of the fricative (C-off), using a 25 ms Gaussian window and a 500 Hz to 10,000 Hz pass Hann filter. The windows were either aligned to fricative edges (C-on and C-off) or centred at the midpoint (C-mid). The low cutoff was set to exclude low-frequency room noise or voicing leakage from surrounding vowels (cf. Zsiga, 2000; Nowak, 2006).

Formants F1, F2, and F3 (Hz) measured at 3 points within the following vowel (or the preceding vowel for word-final fricatives): onset (V-on), midpoint (V-mid), and offset (V-off), using a 25 ms Gaussian window and the Formant (Burg) algorithm. The windows were either aligned to fricative edges (V-on and V-off) or centred at the midpoint (V-mid).

3.2. Cepstral measures

The first 6 cepstral coefficients (c0-c5), Bark-scaled, were extracted from 10-ms frames inside each segment (fricative and adjacent vowels). HMMs were used to divide the segments into regions of internally minimized variance (Viterbi, 1967). Each HMM consists of three states arranged linearly. Each state models one region of a phoneme, and the state's parameters comprise the means and variances of the feature vectors within that region. Each cepstral coefficient was averaged by region. Only the vocalic region adjacent to the fricative was used. The addition of vocalic data constitutes a notable difference compared to the previous analysis by Spinu and Lilley (2016).

3.3. Statistical analysis

For cepstral measures, the means of the features over all of the vectors in each region were calculated and used as input to the statistical analyses. This resulted in 24 measures for each parameter set: 6 coefficients × 4 regions (3 consonantal regions + 1 vocalic region). Following McMurray and Jongman (2011), we conducted multinomial logistic regression analyses with consonant identity (s, sʲ, ʂ, ʃʲ) as the dependent variable and the 24 measures as continuous explanatory variables. For spectral measures, consonant identity was used as the dependent variable and the 16 measures extracted as continuous explanatory variables. Logistic regression has been claimed to be best fitted for categorical response data, as theoretical problems arise with the application of discriminant analyses to them (Morrison and Kondaurova, p. 2160). We used Matlab R2013a (MATLAB, 2013) to determine the regression coefficients. The first analyses were run on the entire data, including both genders. We then separated the corpus into male-only and female-only subcorpora and reran the analyses. We ran additional analyses on the Top 16 predictors from the cepstral set (in order to make it more comparable to the spectral set in terms of number of predictors), and the Top 5 predictors from both sets.

3.4. Results

Figure 1 shows the correct classification rates obtained with the cepstral and spectral set, using (1) all predictors, (2) the Top 16 predictors only, and (3) the Top 5 predictors only. These are shown for males and females combined, as well as for each gender separately. The set of the Top 5 predictors varied somewhat depending on the corpus used, specifically males, females, or males and females combined. For the latter corpus (both genders combined), the Top 5 spectral measures were F2-onset, COG-mid, C-intensity, COG-onset, and F3-onset. The Top 5 cepstral coefficients for the same set were C1.1.vowel, C4.3.fricative, C0.3.fricative, C2.1.vowel, C3.3.fricative (where C# stands for the coefficient, and the middle number stands for the region from which it was extracted). Note that for the spectral method, the 'all measures'

[1] According to Hayes and Steriade (2004), facts about phonetic difficulty are accessible through experiment, vocal tract modeling, and descriptive phonological work. For example, one of the factors that can create phonetic difficulty in obstruents is place of articulation. Ohala and Riordan (1979) found that the size of the cavity behind the oral constriction affects the aerodynamics of voicing, with the time interval from the onset of stop closure to the point where passive devoicing sets in varying with the site of the oral constriction. This asymmetry is found in phonological patterns involving single stops and geminates in many languages: [g] implies [d] which implies [b].